Genomic selection is a promising approach to incorporating markers into breeding programs in order to expedite genetic gain for highly complex traits. Over the last two years, I have been assisting the University of Minnesota Barley Breeding Program implement genomic selection. A pilot study indicated that accurate predictions could be made using relatively small training populations (~300) and marker sets (384). One cycle of selection has shown promising results with respect to gain in FHB severity. A second cycle of selection was just made this last December, and validation trials are being conducted each summer. This program, as well as other breeding programs looking to implement genomic selection, has spurred questions regarding resource allocation in a genomic selection program. Because replication occurs at the allelic level, replication number (r) can be sacrificed in favor of larger population sizes (N). In the context of QTL mapping and traditional marker-assisted selection, it has been shown that increasing N provides more power than increasing r under limited resources. This issue has not been directly addressed for genomic selection. Genomic selection uses all marker information and therefore captures much more genetic variation across many small effect loci, potentially influencing the tradeoff between N and r. Preliminary results using biparental simulations show that enhanced prediction accuracy can be obtained through increasing either r or N. Greater genetic gain is obtained with higher N and thus higher selection intensity and probability of capturing superior recombinants, not through enhanced prediction accuracy.
Marker-Based Selection on Steroids

Optimization of Genomic Selection for Plant Breeding

Aaron Lorenz

UNIVERSITY OF NEBRASKA Lincoln
Outline

• Introduction to genomic selection
• Genomic selection for FHB resistance in barley
  – Review of optimization findings
  – Implementation of genomic selection
• Resource allocation in a genomic selection program
  – Tradeoff between N and r
  – Effect on accuracy and genetic gain
1) Perform selection during off season
2) Increase selection accuracy
3) Increase selection intensity
Works well for simple traits and introgression of major QTL

Harjes et al. (2008)

D. Mackill
*Sub1A* in rice

<table>
<thead>
<tr>
<th>Carotenoids</th>
<th>β-carotene</th>
</tr>
</thead>
<tbody>
<tr>
<td>37.36 µg/g</td>
<td>4.19 µg/g</td>
</tr>
</tbody>
</table>

Wheat Ug99 stem rust

Zhang et al., 2010

<table>
<thead>
<tr>
<th>Carotenoids</th>
<th>β-carotene</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.48 µg/g</td>
<td>5.93 µg/g</td>
</tr>
</tbody>
</table>
Complex traits are controlled by many small-effect alleles

Fat percentage – relatively simple trait

Overall type – complex trait

Hayes et al. (2010)
Role of markers in crop improvement varies

Bernardo, 2008
“Genomic selection” comprises methods that use genotypic data across the whole genome to predict complex traits with an accuracy sufficient to allow selection on that prediction alone.
The essentials

• Genomic selection = Genome-wide selection
  – A form of marker-based selection
    • Avoids QTL mapping
    • Includes all markers in model
      – Unlike “marker-assisted recurrent selection”

• Made possible by:
  – High-throughput marker technologies
  – Relatively new statistical methods
  – High-performance computing

• Advantages over traditional marker-based selection
  – Higher marker density allows marker-linkage phases to be common across broader population
  – Captures small allelic effects
  – No arbitrary statistical threshold
Training population

<table>
<thead>
<tr>
<th>Line</th>
<th>Yield</th>
<th>Mrk 1</th>
<th>Mrk 2</th>
<th>...</th>
<th>Mrk p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line 1</td>
<td>76</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Line 2</td>
<td>56</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Line 3</td>
<td>45</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Line 4</td>
<td>67</td>
<td>0</td>
<td>1</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Line n</td>
<td>22</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

Selection candidates

<table>
<thead>
<tr>
<th>Line</th>
<th>Yield</th>
<th>Mrk 1</th>
<th>Mrk 2</th>
<th>...</th>
<th>Mrk p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line A</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Line B</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Line C</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Line D</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Line n</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

Model training

\[ y_i = \sum_{j=1}^{p} b_j x_{i,j} \]

Prediction

\[ GEBV_i = \sum_{j=1}^{p} \hat{b}_j x_{i,j} \]

Parent selection

<table>
<thead>
<tr>
<th>Line</th>
<th>GEBV</th>
<th>Mrk 1</th>
<th>Mrk 2</th>
<th>...</th>
<th>Mrk p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line A</td>
<td>80</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Line B</td>
<td>67</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Line C</td>
<td>56</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Line D</td>
<td>89</td>
<td>0</td>
<td>1</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Line n</td>
<td>23</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

Basic framework

GEBV = genomic estimated breeding value
The Genomic Selection Cycle

- ** Phenotype (lines have already been genotyped)**
  - Model Training Cycle
    - Advance lines informative for model improvement
    - Train prediction model
  - Updated Model
  - Genomic Selection
    - Advance lines with highest GEBV
    - Genotype selection candidates
  - Line Development Cycle
    - Make crosses and advance generations
    - New Germplasm
  - Test varieties and release

How have we used markers for complex traits in the era of low marker density?

- **Marker-assisted recurrent selection**
  1. Make cross to form biparental population of progenies
  2. Phenotype and genotype
  3. Select markers and estimate marker effects within biparental population
  4. Select and recombine
  5. Repeat
What is the best use of marker information in the era of high marker density?

1. Combine progenies from entire breeding population to estimate marker effects
   • Exploit population-wide LD
   • Minimize phenotyping
   • Estimate allelic effects in context of entire breeding population

2. Use all marker information to capture small allelic effects.
Choice of markers in traditional marker-assisted selection is arbitrary.
Estimation methods for genomic selection

1. Shrinkage models
   • RR-BLUP, BayesA

2. Dimension reduction methods
   • Partial least squares
   • Principal component regression

3. Variable selection models
   • BayesB, BayesC$\pi$, BayesD$\pi$

4. Kernel and machine learning methods
   • Support vector machine regression

<table>
<thead>
<tr>
<th></th>
<th>Yield</th>
<th>Mrk 1</th>
<th>Mrk 2</th>
<th>…</th>
<th>Mrk $p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line 1</td>
<td>76</td>
<td>1</td>
<td>1</td>
<td>…</td>
<td>1</td>
</tr>
<tr>
<td>Line 2</td>
<td>56</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Line 3</td>
<td>45</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Line 4</td>
<td>67</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>…</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Line $n$</td>
<td>22</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Basic genomic selection model

\[ y_i = \sum_j x_{ij} \beta_j + \sum_k w_{ik} u_k + \sum_l z_{il} a_l \delta_l + e_i \]

- Fixed effects for environmental effects and subpopulation membership
- SNP allele effects (random)

\[ z_{il} = \text{allelic state of marker } l \text{ in individual } i \]
\[ a_l = \text{random effect of marker } l \]
\[ \delta_l = \text{indicator variable for the inclusion or exclusion of marker } l \]
Differences between models

RR-BLUP \( a_l \sim N(0, \sigma^2_a) \)

Bayesian \( a_l \sim N(0, \sigma^2_{a_l}) \)

LASSO

BayesB \( a_l = \begin{cases} 0 & \text{with prob } \pi \\ \sim N(0, \sigma^2_{a_l}) & \text{with prob } (1-\pi) \end{cases} \)

BayesC\( \pi \) \( a_l = \begin{cases} 0 & \text{with prob } \pi \\ \sim N(0, \sigma^2_a) & \text{with prob } (1-\pi) \end{cases} \)
Simulated marker effects

Large effect QTL simulated
Statistical methods: assumptions

- **Shrinkage models**
  - RR-BLUP, BayesA
- **Dimension reduction methods**
  - Partial least squares
- **Variable selection models**
  - BayesB, BayesC\(\pi\)
- **Kernel and machine learning methods**
  - Support vector machine regression

- **Large number QTL, equal effects**
- **Few QTL, QTL effects vary**
- **Complex interactions**
GENOMIC SELECTION FOR FHB RESISTANCE IN BARLEY

Credit: Photos from barleyworld.org
Breeding for FHB resistance in barley

- F$_{4.5}$ plant rows
- 2 reps x 2 locs (1.5 m single row plots)
- Mist irrigated and pathogen inoculated nursery
- Visual assessment of 10 heads per plot
- Harvest grain from selected rows
- Gas chromatography analysis of grain for DON (mycotoxin) concentration.
Breeding for FHB resistance in barley

Cost comparison

Rate FHB
$10 per plot
→ $40 for a 2-loc, 2-rep mean

Measure DON on 20% of samples
$10 per grain sample

Genotype with 384 SNP array
$17 per sample
Fusarium head blight resistance data from the Barley CAP project

- Six-row barley germplasm obtained from three Upper Midwest breeding programs
  - University of Minnesota
  - North Dakota State University
  - BuschAg

- 1023 polymorphic markers

- Two related traits
  - FHB resistance
  - DON (mycotoxin) concentration

- 691 barley lines, multiple environments
  - 4 locations per year, 2 reps
  - 224 lines evaluated in 2006
  - 241 different lines evaluated in 2007
  - 226 different lines evaluated in 2008
Exploring genomic selection with the Barley CAP datasets

Training set

CAP 1
2006
n = 224

CAP 2
2007
n = 241

Build prediction models using different statistical methods:
- RR-BLUP
- BayesCπ
- Bayesian LASSO (BLR)

Validation set

CAP 3
2008
n = 226
Summary of findings

• No differences in prediction accuracy between models
• No added benefit to large number of markers
  – 384 worked just fine
• 200-300 is a good training population size
• No apparent benefit to combining subpopulations
  – I.e., if goal is to predict MN lines, MN-only TP just as good as larger TP with both MN and ND lines.
Implementing Genomic Selection, Kevin Smith

Current Scheme

Crossing
A x B
→
F1

Inbreeding
F2
→
F3
→
F4 Winter Nursery
→
F4:5
Plant Rows

Evaluation
Winter Nursery Increase
1st year trial
2nd year trial

AMBA
Pilot Malt
Plant Brew

RELEASE

Breeding Cycle

Genomic Selection

Crossing
A x B
→
F1

Inbreeding
F2
→
F3

F4 Winter Nursery

Evaluation
1st year trial
2nd year trial

AMBA
Pilot Malt
Plant Brew

3rd year trial
4th year trial
5th year trial
6th year trial

RELEASE

Genomic Selection

Yr 1

Yr 2

Yr 3-4

Yr 5-9

Yr 1

Yr 2

Yr 3

Yr 4-8
FHB Genomic Selection Project

Parents

UM | Busch Ag | NDSU

Crosses

UM | B A | B A | NDSU | UM | NDSU

1440 progeny = 6 cross types x 10 crosses per type x 24 progeny per cross

384 SNPs selected from BOPA1&2 optimized for PIC and genome distribution

Training Data set CAPI, II, II from three programs

Ridge Regression Model

Shiaoman Chao, Jean-Luc Jannink, Aaron Lorenz, Rich Horsley, Blake Cooper, Gary Hanning,

Implementing Genomic Selection, Kevin Smith
Evaluation of Cycle 1

Fall 2009  Crosses among parents (UM, BA, NDSU)

Winter 2009/10  F1

Summer 2010  F2

Fall 2010  F3 GS (384 SNP markers)

Winter 2010/11  F4 New Zealand Off Season Nursery

Summer 2011  Yield Trials, FHB Screening, Malting Quality
R = 0.54
FHB Selection

Random N=300
Mn= 1.43

Genomic Selection N=72
Mn= 1.10

Phenotypic Selection N=96
Mn= 1.63

FHB Severity as a proportion of common check QUEST
2011 Trial Means from Crookston and St. Paul

Selected from 1440 progeny from 60 crosses

Parallel phenotypic selection program selected from 1610 progeny from 23 crosses
Distribution of predictions

FHB

Cycle 1

Cycle 2

DON

Cycle 1

Cycle 2
Resource Allocation in a Genomic Selection Program

• Unit of evaluation in phenotypic selection program
  ➢ The individual

• Unit of evaluation in genomic selection program
  ➢ The allele
To rep or not to rep
QTL Mapping

Knapp and Bridges (1990)

Schon et al. (2004)

Figure 4 in original paper

Figure 1 in original paper
Simulation design

- DH Pop 100 QTL
- Train GS model
- Evaluate, Genotype plot $h^2 = 0.20$
- Select 20
- Intermate selected DHs
- Recombine F1s
- Predict genetic value with GS model
- Select 20
- Recombine (Random mate)
- 200 $C_n$ progenies

Only vary population size and experimental design (i.e, rep #) for model training phase

Repeat to Cycle 3
### Budget = 500 plots

<table>
<thead>
<tr>
<th>N</th>
<th>Locs</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>1</td>
</tr>
<tr>
<td>167</td>
<td>2</td>
</tr>
<tr>
<td>125</td>
<td>3</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
</tr>
<tr>
<td>83</td>
<td>5</td>
</tr>
<tr>
<td>71</td>
<td>6</td>
</tr>
<tr>
<td>63</td>
<td>7</td>
</tr>
<tr>
<td>56</td>
<td>8</td>
</tr>
<tr>
<td>50</td>
<td>9</td>
</tr>
<tr>
<td>45</td>
<td>10</td>
</tr>
</tbody>
</table>

Locations serve as reps (i.e., 1 rep per loc)

Genotyping cost per line = 1 plot

\[ B = N + L \times N \]

Genotyping cost

Phenotyping cost
Genetic gain from different GS experimental designs
Selection theory

Response to selection = \( \mu_{r_A} \sigma_A \)

\( r_A = \) selection accuracy = \( r \) (EBV:TBV)

EBV – estimated breeding value
TBV – true breeding value

GEBV – genomic estimated BV

\( r(GEBV:TBV) \) critical metric in genomic selection studies
Simulation design

DH Pop
100 QTL

Train GS model

Evaluate, Genotype plot $h^2 = 0.20$

Select 20

Intermate selected DHs
Recombine F1s

Predict genetic value with GS model

200 $C_n$ progenies

Recombine (Random mate)

Select 20

Repeat to Cycle 3
Increasing L

Increasing N
Prediction accuracy with varying N and L

![Graph showing prediction accuracy with varying N and L]
Conclusions

• Optimistic on genomic selection in UMN Barley Breeding Program.

• Resource allocation
  – Very preliminary results indicate larger N provides greater genetic gain through greater selection intensity.
  – Prediction accuracy can be enhanced by increasing either N or within budget framework.
Acknowledgements

• **Kevin Smith; U of MN**
  – Barley Breeding and Genetics

• **Jean-Luc Jannink; USDA-ARS, Cornell**
  – Small Grains Quantitative Genetics

• **Jode Edwards; USDA-ARS, IA State**
  – Maize quantitative genetics

• Funding
  – USDA Barley Coordinated Agricultural Project
  – U.S. Wheat and Barley Scab Initiative